DOCUMENT RESUME

ED 163 030 TM 008 022

AUTHOR Jackson, Gregg B.
TITLE Meta-Research Methodology.
PUB DATE Feb 78
NOTE 28p.; Paper presented at the Annual Meeting of the American Educational Research Association (62nd, Toronto, Ontario, Canada, March 27-31, 1978)

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS *Behavioral Science Research; Content Analysis; Critical Reading; *Literature Reviews; *Research Methodology; *Research Problems; *Research Reviews (Publications); *Sampling; Social Science Research; Surveys; Technical Reports
IDENTIFIERS Meta Evaluation

ABSTRACT

It appears that relatively little thought has been given to the methods for reviewing, synthesizing, and reporting the results of a set of empirical studies on a given substantive topic. The question is raised as to whether the studies that have been reviewed constitute a population or a sample, and if a sample, whether they are representative or biased or random so that inferences can be tested with confidence. The purposes of this study were to: describe the various methods used for specified aspects of reviews and syntheses, determine the frequency of the alternative methods, critically evaluate the strengths and weaknesses of these methods, and suggest ways in which more powerful and valid reviews and syntheses can be done. The primary source of data was a content analysis of a random sample of reviews and of a sample of allegedly exemplary reviews. The author concludes that although such reviews are important to science and social policy-making, many integrative reviews are done less rigorously than is currently possible. This paper and the recent work of Gene Glass suggest several ideas for improving the prevailing methods for reviews. (Author/CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

25,12

ED163030

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Gregg Jackson

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

META-RESEARCH METHODOLOGY

Gregg B. Jackson

Social Research Group
The George Washington University
2401 Virginia Ave. NW
Washington, D. C. 20037

February 1978

For presentation and discussion at
the annual meeting of the
American Educational Research Association
Toronto, Canada, March 27-31, 1978

This paper is a brief report of research supported by grant #DIS 76-20398 from the National Science Foundation. A much more detailed report of the work is provided in the final project report (Methods for Reviewing and Integrating Research in the Social Sciences, by Gregg B. Jackson) which will be available through the National Technical Information Service by June of 1978.

Many persons have provided assistance with this research, most importantly Jomills Braddock, Ira Cisin, and Judith Miller.

TM008 022

Introduction

Reviews and syntheses of empirical research studies on a given topic are a fundamental activity in behavioral research; they usually precede any major new research study, and also are done as independent scholarly works. This paper reports a recent investigation of the methods used for such reviews. The investigation was limited to reviews that are focused on making inferences about substantive issues from empirical research. These will be called integrative reviews in this report. Excluded were reviews of theoretical positions, of methods, and of non-empirical research.

Given the importance and widespread conduct of integrative reviews, one might expect that there would be a fairly well developed literature on methods, techniques, and procedures for conducting such reviews; but this is not the case. An earlier examination by this author of a convenience sample of 39 books on general methodology in sociological, psychological, and educational research revealed there was very little explanation of matters other than the use of card catalogs, indexes to periodicals and note-taking. Only four of these books discussed how to define or sample the universe of sources to be reviewed, three discussed criteria by which to judge the adequacy of each study, and only two discussed how to synthesize validly the results of different studies. None of the discussions exceeded two pages in length.

Similarly, a preliminary examination of journal article titles in Sociological Abstracts (from January 1973 through October 1975), Psychological Abstracts (from January 1973 through December 1975) and Current Index to Journals in Education (from January 1973 through June 1975) revealed a dearth of work on integrative review methods. Entries under the following subject headings were examined: "literature reviews," "methods," "methodology," "research methods," and "research reviews." Only five of the titles out of approximately 2,050 entries appeared directly relevant. Upon examination, one of the sources proved to be inappropriate and another could not be located. The remaining three will be discussed briefly later in this section.

Additional evidence that there are few explicated methods, techniques and procedures for integrative reviews is the fact that few published integrative reviews adequately describe the methods used. A preliminary examination of 87 review articles in the 1974 and 1975 volumes of American Sociological Review, Sociological Quarterly, Social Problems, Psychological Bulletin, and Review of Educational Research found only twelve articles which provided some statement on the methods used.

Doing a good integrative review is never easy. It might seem that when all or almost all of the studies on the topic yielded similar results, the work would be easy, but this is incorrect because a careful reviewer is still obliged to determine whether all the studies have biases in the same direction which caused similar but invalid results. In the

more prevalent case where the studies on the topic have different, and apparently contradictory, results, the work is obviously difficult. A good review of such research should explore the reasons for the differences in the results and determine what the body of research, taken as a whole, reveals and does not reveal about the topic.

The most valuable previous writings on integrative review methods have been done during the last decade.

Kenneth Feldman (1971) wrote that there is "...little formal or systematic analysis of either the methodology or the importance of... reviewing and integrating...the 'literature'..." (p. 86). He suggests that, "half-hearted commitment in this area might account in part for the relatively unimpressive degree of cumulative knowledge in many fields of the behavioral sciences" (p. 86). He mentioned the problem of not being able to know the parameters of the universe of relevant studies. Feldman suggested the utility of examining the distributions of results in more than one manner, and suggested that inconsistent results can sometimes be explained by differences in subjects, treatments, settings, and the quality of the research methods. He warned reviews should avoid hypercriticalness as well as hypocriticalness, and indicated that a good review of research "shows how much is known in an area, [and] also shows how little is known" (p. 100).

Richard Light and Paul Smith's excellent article (1971) discussed the present lack of systematic efforts to accumulate information from a set of disparate studies. Light and Smith used a four category typology to characterize most present integrative reviews. The first category comprises those reviews which merely <u>list</u> any factor which has shown an effect on a given dependent variable in at least one study. A second category comprises reviews which <u>exclude</u> all studies except those which support one given point of view. The third category is for those reviews which, in one way or another, <u>average</u> the relevant statistics across a complete set of studies. The fourth category is comprised of <u>vote taking</u>—counting the positive significant results, the non-significant results, and the negative significant results, and if a plurality of studies have one of these findings, then that finding is declared the truth.

Light and Smith pointed out the weakness and resulting consequences of these procedures, and proposed as a superior alternative a paradigm for secondary analysis of data from various studies which have a common focus. The paradigm suggests the data ought to be analyzed <u>within</u> strata that take into account different characteristics of subjects, treatments, contextual variables, and interaction effects among these. Ironically, Light and Smith failed to point out that such a paradigm could also be useful for integrating results of different studies when secondary data analysis is not feasible. (Time constraints, promises about the confidentiality of data, lost data sets and other factors sometimes can preclude secondary data analysis.)

Gene Glass (1976) presented an important paper on what he called "meta-analysis of research." After stating the need for "a rigorous alternative to the casual, narrative discussions of research studies which typify our attempts to make sense of the rapidly expanding research literature," Glass proposed such an alternative. He suggested expressing the results of studies on a given topic in a common metric, coding the various characteristics of studies that might have affected their results, and then using multiple regression equations or other statistical techniques to study the association of variations of those characteristics with the variations in the results. This approach differs from secondary data analysis in that it does not use the data on the individual subjects within one or more studies, but rather uses data on the overall characteristics of each study.

The lack of explicit methods for doing integrative reviews is a serious problem for at least four reasons. First, the lack of explicit methods appears to be in large part the result of social scientists failing to give much thought to such methods, and thus it probably means that they are not using as powerful methods as could be developed for accumulating social science evidence. Second, it makes it difficult to have standards for judging the quality of integrative reviews. Third, it makes it difficult to train graduate students to do competent research reviews. Fourth, the lack of review methods hinders the accumulation of valid knowledge from previous research.

Despite the lack of explicit methodology for doing integrative reviews, each review is the result of implicit methods, consciously or unconsciously selected by the reviewer.

This study primarily focused on the methods that are currently being used for integrative reviews of empirical research in sociology, psychology and educational research. The study had four objectives:

1. To develop a conceptualization of the various methodological tasks of integrative reviews and of the alternative approaches to each task;
2. To estimate the frequency with which current reviews published in high quality social science journals used each of the alternative approaches;
3. To evaluate critically the strengths and weaknesses of the alternative approaches;
4. To suggest some ways in which more powerful and valid integrative reviews might be done.

Data Collection

This project investigated several sources of information on the methods used for integrative reviews. These sources were:

1) a purposive sample of 16 integrative review articles that were suggested by various persons as being methodologically exemplary;

2) a random sample of 36 integrative review articles from 1974, '75 and '76 volumes of prestigious social science periodicals (Psychological Bulletin, Annual Review of Psychology, Review of Education Research, Review of Research in Education, American Sociological Review, Sociological Quarterly, Social Problems, and Annual Review of Sociology); the sample was stratified so as to yield 12 articles from each of the three disciplines and equal numbers of articles from each source with a discipline;

3) published rejoinders to the 36 randomly sampled articles;

4) Gene Glass's recent papers on or using meta-analysis (1976a, 1976b, 1977a, 1977c, Smith & Glass, 1977) and personal communications with him;

5) responses to queries sent to editors of prestigious social science periodicals that frequently publish integrative reviews (Psychological Bulletin, Annual Review of Psychology, Annual Review of Sociology, Review of Research in Education, and Review of Educational Research);

6) responses to queries sent officials of national organizations that were thought to have major responsibilities for reviewing and synthesizing research in the social, biological or physical sciences (Consumer Information Branch, National Institute of Education; Assembly of Behavioral and Social Sciences, National Academy of Sciences; Congressional Reference Service, Library of Congress; Information Systems Operation Branch, National Institute of Education; Special Studies Division, Office of Research and Development, Environmental Protection Agency; Developmental Neurology Branch, National Institute of Neurological and Communicative Disorders and Stroke; Office of Space Science, National Aeronautics and Space Administration; Program Analyses and Formulation Branch, National Cancer Institute; Assembly of Life Sciences, National Academy of Sciences; Assembly of Mathematical and Physical Sciences, National Academy of Sciences.

This paper will focus primarily on the results from the coding and analysis of the 36 randomly sampled articles and on a brief critique of Glass's proposed meta-analysis.

The purposive sample of allegedly methodologically exemplary review articles was examined primarily to aid in conceptualizing the nature of

review methodology and to suggest desirable approaches for various methodological tasks of a review. Following an examination of the allegedly exemplary articles, and an examination of a substantial number of other review articles, it was decided to conceptualize the methodology of integrative reviews as involving seven basic tasks: 1) selecting the topic(s), 2) consulting previous reviews on the same topic or similar topics, 3) sampling the research studies that are to be reviewed, 4) representing the characteristics of the studies and their findings, 5) analyzing the characteristics of the studies and their findings, 6) interpreting the results and 7) reporting the review.

These tasks are analogous to those engaged in when doing primary research (research that involves collecting original data on individual subjects or cases). Indeed, this conceptualization was based on the presumption that reviewers and primary researchers share a common goal and encounter similar difficulties. The common goal is to make accurate generalizations about phenomena from limited information.

Since the methodology of primary research was used to conceptualize the methodology of integrative reviews, the standards for competent primary research were thought to be the appropriate evaluative criteria for judging the alternative methodological approaches that were investigated in this study. The problem, of course, was to decide what those standards are. Though the methodology of primary research is more highly developed than the methodology of integrative reviews, there are many aspects of it about which there is much disagreement among social scientists. There is much agreement, however, on certain topics. Sampling theory, as discussed, for instance, by Kish (1965), is widely thought to provide the best guidelines for samples when the purpose is to make generalizations from a relatively small sample to a much larger population. There is much agreement (and some disagreement) on the appropriateness of alternating descriptive and inferential statistics that are described, for instance, by Bradley (1968), Hays (1963), and Kerlinger and Pedhazur (1973). And there is substantial agreement on some of the major threats to the internal and external validity of any study, as discussed, for instance, by Campbell and Stanley (1963) and Bracht and Glass (1968).

A coding instrument was developed to code various aspects of, and approaches to, each of these tasks. The final version of the coding instrument had 66 items. It was used to code the 36 randomly sampled articles. Each article was coded independently by two coders, and all discrepant codings were afterwards resolved by the coders. Intercoder reliability and reliability over time were assessed and found quite satisfactory.

The Results and Discussion of Random Sample

The instrument used to code the random sample of integrative review articles is 15 pages long and is not appended to this paper. The numbers preceded by a V and enclosed in parentheses in various places throughout the following text refer to the coding instrument item numbers; copies of the instrument or clarification of specific items are available from the author.

Only point estimates are reported in the text. Most, but not all of them, are based on 36 units of analysis. A 0.95 confidence interval for $N = 36$ and $X = 6$ would be $1 \leq \hat{X} \leq 12$; for $X = 12$ it would be $7 \leq \hat{X} \leq 19$; and for $X = 18$ it would be $11 \leq \hat{X} \leq 25$. A 0.95 confidence interval for $N = 20$ and $X = 5$ would be $1 \leq \hat{X} \leq 10$; and for $X = 10$ it would be $5 \leq \hat{X} \leq 15$.

Task 1: Selecting the Topic(s)

Data were not coded on this task since it was not recognized as an important part of doing a review until the coding had been completed.

Task 2: Use of Previous Reviews (V11 and 12)

Just as reviews of past primary research are useful for preparing and interpreting present research on a given topic, previous reviews of a topic can be quite useful for preparing and interpreting present reviews of research literature. Judging by the frequent citation of earlier reviews in the sample of review articles, there seems to be rather widespread agreement on this point. Seventy-five percent of the 36 randomly sampled integrative review articles cited previous reviews on the topic or on similar topics (V11), but only two of these 27 provided any critique of the previous reviews (V12).

The uncritical acceptance and use of previous reviews is as undesirable as the uncritical acceptance and use of any other research. One of the widely recognized responsibilities of a researcher is to examine critically all evidence used in his or her research. Important decisions about the focus, methods, and interpretations of a cumulative review are probably sometimes heavily influenced (consciously or unconsciously) by examinations of previous reviews on the topic or similar topics. There is nothing incorrect or undesirable about this if the reviewer uses scholarly judgment in evaluating the strengths and weaknesses of the previous reviews.

A good example of critically examining previous reviews on the topic is provided by Lambert (1976). He examined a number of previous reviews on his topic, sought reasons for their discrepant conclusions, and then used that information to improve the procedures of his own review.

Task 3: Sampling (V16, 19, 24a-c, and 45)

The results of any integrative review will be affected importantly by the population of primary studies that it is focused upon and by the manner in which the actually reviewed studies are selected from that population.

Only one of the 36 randomly selected review articles reported using indexes (such as Psychological Abstracts of Dissertation Abstracts) or information retrieval systems to locate primary or secondary studies for possible inclusion in the review (V16). Only three of the 36 review articles reported searching bibliographies of previous reviews or querying experts on the topic in an effort to locate appropriate sources for their review (V19).

It seems reasonable to assume that these results mainly reflect reviewers' failure to report how they searched for sources, rather than a failure to use the indicated means for the search. It is almost inconceivable that most reviewers do not use indexes or bibliographies. The failure of almost all integrative review articles to give information indicating the thoroughness of the search for appropriate primary sources does, however, suggest that neither the reviewers nor their editors attach a great deal of importance to such thoroughness.

The question of whether a set of located studies on a topic ought to be considered a sample or a population is a difficult one. But in either case it is highly desirable to locate as many of the existing studies on the topic as is possible. Since there is no way of ascertaining whether the set of located studies is representative of the full set of existing studies on the topic, the best protection against an unrepresentative set is to locate as many of the existing studies as is possible. Then if the number of located studies is greater than can be carefully reviewed, a sample of those studies can be used in the review.

Data were collected on the extent to which the reviewer discussed or analyzed the full set of located studies on the topic. One of the 36 randomly sample reviews discussed or analyzed the full set of located studies, 6 of the reviews clearly did not, and for the 29 other reviews, the information given in the published article was insufficient for making a judgment on this matter (V45).

Data were not collected on how reviewers selected studies for analysis or discussion from the located studies. The coders' impression, however, from reading the reviews, is that subsets were usually purposive samples of "methodologically adequate studies" or of "representative" studies. For instance, Glass (1976a) analyzed only those studies that had a control group. Sechrest (1976) indicated, "An annual review, even in an area so circumscribed as personality, cannot serve as a substitute for Psychological Abstracts. No pretense or breadth or depth of coverage is made here. The materials cited were chosen because they fit a topic or illustrate a point to be made" (p. 9). And Demerath and Roof (1976) indicated, "It is manifestly impossible to summarize the entire recent literature. Instead, we have highlighted empirical studies that mark significant conceptual and/or methodological advances" (pp. 19-20).

For the purpose of making generalizations, some sort of random sample (simple, stratified, multiphase, etc.) is the most appropriate. Such samples do not assure a representative sample, but neither does any other approach, and random samples have the advantage of allowing an estimate of the probability of drawing a significantly unrepresentative sample.

It is possible for a review to analyze more than one subset of located studies on the topic. Sometimes there may be justification for including an examination of a purposively sampled subset that is exemplary in some manner, but since the definition of an integrative review used in this study is limited to those in which generalizations are sought, there does not seem to be any justification for using a purposive sample except in conjunction with a random one.

Task 4: Representing Characteristics of the Primary Studies (V53a-e and 60)

The representation of the characteristics of the primary studies is, in effect, the data collection of integrative reviews. The manner in which this is done can substantially affect the results and interpretation of the cumulative review.

Twenty-eight of the 36 randomly sampled reviews either reported the findings of many of the individual reviewed studies or indicated how many or what percentage of the studies had each type of finding or result (V44a). Of the 28, eighteen represented at least one finding of the primary or secondary research with an indication of the direction and magnitude of the difference or of the association (by standard score difference measures, or r_p, r_s, T, R, W, W^2, r_p^2, R^2, etc.) (V53a). Only 4 of the 28 reviews made at least one clear distinction among: significant positive findings, non-significant positive findings, non-significant negative findings, and significant negative findings (V53b). Only one of the reviews clearly represented the findings of the primary studies in any of the other investigated ways (V53c-e). For 10 of the 28 studies, there was insufficient information for judging how the reviewer represented any of the findings of the primary studies. It should be noted that information on items V53a-e was coded as Yes if there was any instance in the review that reflected the item. Consequently, it was quite possible for a review to be coded as having represented a finding of a primary study with a magnitude measure, and yet for there to have been no clear indication of how the reviewer represented most of the findings. It was also possible and fairly common for the reviewer to report the findings of many individual studies, but in a manner such that it was impossible to judge how the reviewer had represented most of the findings. The impression of this writer is that such ambiguities were present in about 80 percent of the review articles. It was common to find reports that "Johnson found a relation between X and Y, but Alexander and Henderson did not." It was often impossible to know whether reported relations were statistically significant or included those that were "substantially" different from zero but not statistically significant. It also was common for findings of primary studies to be reported as statistically significant with no explicit indication of their direction.

Every reviewer has to represent the findings of the primary studies in some manner, and though items V53a-e may not exhaust the ways in which this can be done, they almost certainly include the ways most commonly used. The alternatives are mentioned above in order of

decreasing amount of information they provide; they descend from an interval measure, to ordinal measures, to a nominal measure. Of the alternatives, the magnitude measures with a directional sign (V53a) are clearly the preferred way of representing the findings of the primary studies. To analyze these it is necessary to reduce them to a common metric, a chore that is not always easy, but one on which some development work is currently being done (Glass, 1977a, 1977c). The next best alternative is representing the findings as significant (+), non-significant (+), zero, non-significant (-) and significant (-) (V53b). This alternative may be best if magnitude measures or the data needed to calculate them are not reported in many of the primary studies, but it is quite inferior to the first approach, as will be discussed in the next subsection. The worst of the alternatives indicated in the coding instrument generally is the last (V53e), where the findings are represented as a significant difference in one given direction or not so. This alternative should usually be avoided, for it produces ambiguous data, unless most of the primary studies being reviewed used one-tailed tests of their hypotheses. For instance, if 12 out of 26 findings are significantly positive, is this good evidence of a positive relation for the studied phenomena? It largely depends on how many of the remaining 14 studies had significantly negative findings which cannot be determined from this representation.

No data were collected on how the reviewer represented the independent variables of the primary studies. A preliminary examination had shown that review articles hardly ever indicate this. The representation of the independent variables of primary studies can have a major impact on the results of a review.

Only one of the 36 randomly sampled review articles indicated that the reviewer, when encountering reports of primary and secondary studies that did not have all the information needed for the analyses, sought to get the information from the authors of the reports of those studies or from detailed final reports of funded research, or calculated or estimated the information from the other information given in the initially reviewed report of the study (V60). It is just about inconceivable that 35 of the 36 reviews did not encounter problems with missing information. What cannot be determined from this study is whether the failure of review articles to report efforts to get such information reflects an omission in the reports or an omission of efforts to get the information. This writer suspects that it is some of each, but predominantly the latter.

In primary research today it is quite common for the investigators to make rather extensive efforts to minimize missing data and to report those efforts briefly. It would appear that similar efforts and reporting procedures are equally desirable for reviews.

Task 5: Analyzing the Primary Studies (V41, 56a-c, 57a-e, and 62)

Analysis is the process by which the reviewer makes inferences from the primary studies. It includes: judgments about the implications of

identified methodological strengths or weaknesses in the primary studies, estimates of population parameters of the studied phenomena, and assessments of how varying characteristics of subjects, content, and treatments or suspected causal variables may affect the phenomena. Twenty-six of the 36 reviewers described what were considered to be the major methodological difficulties or shortcomings of the primary research that was reviewed (V41). Some of the other 10 reviewers may have examined these difficulties or shortcomings but failed to report on them.

If more than a small portion of the reviewed studies have serious methodological weaknesses, these limitations can sometimes lead to invalid inferences unless their effects are considered before drawing inferences about the topic. No data were collected on how identified weaknesses in the methods of the primary studies were taken into account when making inferences from those studies. The impression of this writer is that the most common approach was to indicate that inferences about the topic were unreliable if many weaknesses were found. The second most common approach appeared to be to discard the methodologically "inadequate" studies and base the inferences on the remaining ones. A third approach that appeared to be used in at least a few reviews was to identify weaknesses in the research which supported one point of view and thus discredit the evidence for that point of view, without applying the same standards of methodological adequacy to the research which supported another point of view. All three strategies raise the question of what constitutes a serious threat to the validity of a given study and what does not. There is no simple answer. If a modest number of the studies are devoid of such threats, the impact of the threats in the other studies can be examined empirically in a manner that will be discussed later in this paper.

It should be noted that the actual threat to the internal and external validity of a study is not determined exclusively or even primarily by the design of the study. Campbell and Stanley's important and widely read monograph (1963) on experimental and quasi-experimental designs shows which threats are controlled by various different designs. But the monograph does not indicate which threats are likely to be trivial in a given study nor which threats can be reasonably controlled by other means. For instance, instrument decay may be a serious threat to internal validity when using a rater's judgment of people's emotional health, but is unlikely to be a serious threat when measuring children's height using the kind of device that is common in physicians' offices. Similarly, obtrusive measures of a variable pose a more serious threat of testing effects than do unobtrusive measures. And studies where the data are collected over a brief period of time are less likely to have their validity threatened by history and maturation than are studies where the data collection extends over a longer period of time.

It is the impression of this writer that some reviewers will label as methodologically inadequate any studies which do not have experimental or strong quasi-experimental designs. Sometimes this is appropriate, but the above discussion ought to indicate that it is not always appropriate.

Often, but not always, when there is a sizable number of studies on a given social science topic, there are some results which appear incongruent with the other results. There are a number of possible reasons for varying results in a set of studies on a given topic. One of these is random sampling error. Sampling theory indicates that, when there is a set of studies from a given population, the findings will vary some. About half of the study findings will be greater than the population parameter and about half will be less than the population parameter. In addition, if each study's findings are tested for statistical significance at the .05 level with the null hypothesis being the true population parameter, about 2.5 percent will have findings statistically significantly greater than the population parameter, and about 2.5 percent will have findings statistically significantly less than the population paramater. This sampling error has to be taken into account when judging whether or not variations in the findings should be considered congruent.

There is strong evidence that some of the reviews failed to take this source of variation into account. Barnes & Clauson (1975, p. 651) reported, "The efficacy of advanced organizers has not been established. Of the 32 studies reviewed, 12 reported that advance organizers facilitate learning and 20 reported that they did not." But an examination of the evidence indicates that the 12 studies yielded statistically significant positive findings, and the other 20 comprise studies which yielded non-significant (+) findings, zero difference, non-significant (-) findings and perhaps significant (-) findings. Barnes and Clawson did not report how many of the 20 studies yielded each type of finding. If the population value was zero, and all the 32 studies tested their hypotheses at the .05 level, then it is expected that between zero and two of the studies ($.05 \cdot 1/2 \cdot 32 = .8$) would have statistically significant (+) findings rather than the 12 that actually did. Unless several of the 20 studies had statistically significant (-) findings, Barnes and Clawson's data strongly suggest that advanced organizers have at least a small positive effect on learning. If there are considerably more than the expected number of both significant (+) and significant (-) findings, it is possible that the population has a bimodal distribution, or that the examined studies were of two or more populations, despite appearances to the contrary.

Another example of reviewers failing to take sampling errors into account when making inferences from a set of studies is the review by Schultz and Sherman (1976). Twenty-two of the 62 studies cited in this review had significant (+) findings; no information is provided on the number of significant (-) findings. Schultz and Sherman wrote:

> The many nonsupportive studies, the qualifications to some of the supportive studies, and in particular, the consistent failure to replicate interactions between social class and reinforcers lead us to several conclusions. 1) Social class differences in reinforcer preferences can not be assumed. (p. 39)

If the population value were zero, 62 studies tested at the .05 level would be expected to yield between zero and seven significant (+) findings rather than the 22 that did occur. There is, however, a factor suggested

by Schultz and Sherman that does complicate the interpretation. They claim that methodologically superior replicates of earlier studies that had found significant (+) findings often failed to yield such findings. This does raise a legitimate concern, but there are some questions as to its implications. First, Schultz and Sherman indicated that only three of the 22 studies with significant (+) findings were unsuccessfully replicated (by a total of 9 studies). Second, Schultz and Sherman did not indicate whether these failures to replicate yielded non-significant (+) findings, non-significant (-) findings or significant (-) findings, and such information is important in interpreting the findings. Third, investigators who conduct a replication may be predisposed to disprove the original study, and these predispositions may create some biases in their investigations despite some real methodological improvements.

There were a number of other reviews examined in this study that also disregarded the distribution of the findings of the reviewed studies, but these reviews also suggested a number of reasons why the bulk of their reviewed studies might be invalid and thus provided some justification for the omission.

Some other reviews had so few studies on the topic that it would be impossible to infer reliably whether all but the most skewed distributions could not reasonably be expected to come from a population where there was "no difference." In the random sample of 36 examined reviews, however, there were at least 18 studies which did not provide adequate information for judging whether or not the reviewer had interpreted variations in the findings of the primary studies in light of expected sampling error (V44a and 53a-e).

Care has to be exercised when analyzing the distribution of findings among the four categories of "statistically significant (+)," "non-significant (+)," "non-significant (-)," and "significant (-)." One complication is that the above discussion has to be modified unless the null hypotheses tested in each primary study were ones of "no difference" or "no relationship." Null hypotheses usually are stated as such, but occasionally they are not. A second complication is that the above discussion presumes that all the tests of hypotheses were two-tailed tests. A third complication is that the above discussion presumes that all the primary studies tested their hypotheses at the same level of Type I errors. A fourth complication is that the above indicated method of analysis does not provide information on the magnitude of the differences or relationships. If the sample sizes of many of the primary studies are quite large (say greater than 500), the method would lead to the conclusion that there is a difference even if the population parameter is only trivially greater or less than zero. This conclusion would not be incorrect but it would be unimportant.

It is also possible to analyze the distribution of findings among the two categories of "positive" and "negative." This can be done by using the binomial distribution, or an approximation of it. This method is

subject to all the above-mentioned complications except the third one. It can yield trivial conclusions either if many of the N's of the primary studies are quite large or if a quite large number of findings is analyzed.

In addition to the sampling error, there are at least three other causes of variations in the findings of a set of primary studies on a topic. These include: the studies in the set examined different phenomena, the studies in the set examined the same phenomena under differing circumstances which affected the findings, or the methods of the studies varied and affected their findings. These reasons can be tested by examining the relationships of the varying characteristics of the studies to the varying results. None of the 36 randomly sampled reviews did such an analysis in a multivariate manner where two or more of the varying characteristics of the primary and secondary studies were simultaneously tested for relationships with the varying results (V56a). Two of the 36 reviews did univariate analyses, examining the relationship of a single characteristic of the studies to the varying results (V56b). Another five of the 36 reviews made such analyses in a systematic discursive manner whereby they discussed how one or more characteristics of the primary and secondary studies were related to differences in the findings across the full set of analyzed studies (V56c). A total of only 7 out of the 36 articles reported analyses by any of the above three means.

This is not because the other 29 reviews did not have discrepant results. As was reported earlier, 32 of the 36 reviews had at least some incongruent findings (V24a). Perhaps the 29 reviewers did such analyses, but did not find statistically significant results, then chose not to report the results; or perhaps they simply failed to do such analysis.

It should be noted that the reviews were not coded as using systematic discursive analysis (V56c) unless they discussed how a characteristic of the study related to differences in the findings across all or most all studies in the analyzed set. The impression of this writer is that most of the reviews did suggest some explanation for the observed differences in the findings and many offered some evidence for the explanation, but that evidence was usually less systematic than coded in V56a-c. For instance, a reviewer might point out that the study that had the highest Y also had a higher X than the study that had the lowest Y, while not mentioning the relation between X and Y in the other studies on the topic.

It is not at all clear why systematic analyses of the correlates of varying findings are not done more often in integrative reviews. Perhaps it is because the reviewers find so many differences among studies that they despair of being able to find systematic relations. Or perhaps it is because the reviewers have simply not thought to do such analyses.

Whatever the reasons, the effect of this omission is obvious and serious. Without such analyses reviewers will sometimes incorrectly infer that the findings of a reviewed set of studies are contradictory and that the available evidence is inconclusive. It seems almost certain

that some of the confusion that surrounds many topics in the social sciences is partly a result of reviewers' frequent failure to search systematically for explanations of the varying results. Multicolinearity or weak correlations will sometimes preclude explanations of the variations, but the search ought to be conducted despite such possibilities.

Sometimes when doing a review, one or two primary studies may be located that offer unusual potential for shedding light on some important issue, if only their original data could be reanalyzed. In such cases a secondary analysis is appropriate. Only one of the 36 randomly sampled reviews reported having done a secondary analysis (V62). It seems unlikely that such analyses would be done and not reported, but it is not at all clear how many times there was justification for doing such analyses.

Sometimes secondary data analysis can be done with a minimum of resources, but sometimes it cannot. Data sets are sometimes lost or inadequately documented; in addition, promises of confidentiality sometimes make it impossible to release unaggregated data.

It should be noted that close congruence among the findings of a set of studies on a given topic does not necessarily indicate that the evidence is valid, and the lack of close congruence among the findings does not necessarily indicate that the evidence is inconclusive. For the purposes of this discussion, the findings of a set of studies on a given topic will be considered congruent if they do not vary more than could be expected by chance from random sampling error. It is possible for the findings to be congruent, but to be invalid. This could occur if all the findings were biased by one or more methodological flaws that were common to all of the studies, or if all the findings had the same net bias but caused by different methodological flaws in different studies. The latter is possible, but not particularly likely.

It should also be noted that one or more methodological flaws in a study, even when serious ones, need not cause biased findings. They only create a threat to validity which may or may not cause a bias.

If the findings are incongruent it is still possible for all of them to be valid. This might be so when the varying measures of the outcome variable represent several somewhat different constructs or when subject characteristics, scope conditions or contextual variables vary among the studies and affect the outcome variable. For instance, the relationship between X and Y may vary in different regions of the country, over different age groups of subjects, or under different social, economic or political circumstances. If the different studies varied in respect to these factors, their results might vary substantially and yet all might be perfectly valid. This is why it is important in integrative reviews, when the findings are not congruent, to search for and examine factors which may systematically co-vary with the findings.

Glass's Meta-analysis

Glass's meta-analytic approach involves transforming the findings of individual studies to some common metric, coding various characteristics of the studies, and then using conventional statistical procedures to determine whether there is an overall effect, subsample effects, and relations among characteristics of the studies and the findings. The original data for each unit of analysis in a study are not used. Rather, the unit of analysis is the study, and summary data from each study are analyzed. For instance, if there is a set of experimental studies which investigate the impact of X on Y, for each study one might code the average age and SES of subjects, the duration of treatment (X), the setting in which the treatment was applied, an estimate of the reactivity or fakeability of the outcome measure used, an estimate of the internal validity of the research design, and the date when the study was conducted. Then these variables would be used in an univariate or multivariate manner to predict a standardized measure of the findings. Glass suggests that when most of the studies are experiments with a control group, the standardized measure of the findings be a standard score difference measure calculated by the mean difference of the experimental and control group divided by the within group standard deviation of the control group (Glass, 1977c, p. 39). He suggests that if most of the studies are correlational, and use different measures of association, the standardized measure be a product-moment correlation; he provides formulas for estimating product moment correlations from various other measures of association such as the point-biserial correlation, Spearman's rank-order correlation, Mann-Whitney U, as well as t and F (Glass, 1977a, pp. 4-10).

The meta-analytic approach has a number of strengths. First, it is a systematic, clearly articulated, and replicable approach to integrating results from a set of studies. Second, it can be used with information from both the best and the less-than-best studies on a topic, but with controls for possible biases caused by various flaws in the available studies. Third, it can provide estimates of the population parameters. Fourth, when using multivariate statistical procedure, it provides a method for simultaneously investigating the relationships of variations among studies in respect to their population of subjects, their scope conditions, the intensity and duration of their treatments, and other factors, with variations in the findings. No approach commonly used to date for doing analyses in integrative reviews has been capable of doing this.

Glass has indicated some difficulties and unresolved questions about the application of his approach. He has pointed out: 1) it is sometimes difficult to get a standardized measure of the finding from a study because of insufficient data, 2) variance stabilization transformations may be desirable for measures of the finding where distribution of the criterion variable is attenuated, 3) the normal distribution assumption when transforming dichotomous data via probit transformations needs to be examined, 4) there is a problem of how to analyze results that are nested within variables analyzed in a study, 5) findings perhaps should be weighted by their sample size, and 6) there are problems of analyzing aggregate data (the study as the unit of analysis) when trying to make inferences about unaggregated phenomena.

There are some other limitations and problems in the application of this approach which have not yet been discussed in published form, and which will be mentioned below. It should be noted that most of these difficulties are common to all analytic approaches to integrative reviews. Nevertheless, they are important to keep in mind when doing or interpreting meta analyses.

One limitation of the meta-analytic approach is that it can assess only relatively direct evidence on a given topic. Sometimes a topic of importance has not been directly investigated but there are studies with indirect evidence that can be reviewed and woven together. For instance, if the topic is "Will substance X reduce chronic depression in adults?", there may not have been any studies on that question, but there may have been studies of the effects of X on depression in baboons and studies of the similarity of effects of other chemicals on depression in baboons and humans. The meta-analytic approach can be used for evaluating the results within each set of studies, but it cannot weave together the evidence across sets of studies on related topics.

A second limitation of the meta-analytic approach is that it cannot be used to infer which characteristics of studies on a given topic caused the differing results. Statistical analyses can provide good evidence of causal relations only when the data are from experiments or strong quasi-experiments. The characteristics of reviewed primary studies are not systematically manipulated in an experiment or quasi-experiment, even when all the studies used experimental designs to investigate the given topic.

The third limitation of meta-analysis is applicable when the set of primary studies is a sample from a larger population and when multivariate statistics are used to analyze the findings. Under such circumstances, there must be a substantial number of primary studies on the topic, but there are no clearly documented standards for sample sizes when doing multiple regression. Kerlinger and Pedhazur suggest at least 30 cases for each predictor (1973:282). Other well-respected statisticians think these suggestions are excessive (Coleman, 1975, Glass 1977b). It should be noted, however, that the number of cases may well be greater than the number of studies, because Glass suggested using an "effect" as the unit of analysis in meta-analysis and each study may have more than one "effect." An effect is defined as any analysis within a study of a given treatment and outcome at a given time of measuring the outcome.

A fourth problem when doing meta-analyses is deciding whether or not a set of studies on a topic ought to be considered a universe or a sample. This has a bearing on whether tests of statistical significance are appropriate, and the number of cases needed to use various statistical tools appropriately. Some sets are obviously samples, such as when a random sample of articles is drawn from a specified sampling frame or when a convenience sample is assembled (the latter does not meet the assumptions of inferential statistics). When the set is a result of a thorough search, the matter is not so clear. First, it is quite likely that even a thorough search will miss some, if not many, of the unpublished studies. Second, even if the search was successful in locating virtually all of the completed studies on the topic, these studies

might be considered only a sample of the phenomena being studied or a sample of all possible studies on the topic. Glass initially suggested that the located studies be considered a population (1977b), but he subsequently has treated them as samples (1977a, 1977c). This writer's tentative opinion is that the set of studies should usually be considered a sample because the analysis of an integrative review is usually intended to make inferences about the phenomena investigated in the individual studies rather than about studies on the phenomena.

A fifth problem when doing meta-analysis is the lack of common metrics for the measures used and reported in the various primary studies on the topic. There are at least three aspects of this problem. First, different constructs are sometimes studied under a single topic. For instance, the outcomes of various studies on the effects of psychotherapy include emotional health, happiness, social relations, and others. Second, for any given construct there are alternative measures whose metrics may not be equivalent. For instance, what is described as upper middle SES on one measure may be described as middle SES on a second measure. Third, the statistics used to measure a relationship between two or more variables can vary in different studies. Studies may use r_p, t, rho, tau, or others.

Glass has suggested the first aspect of the problem is often not serious and can be ignored (1977d). He would argue that all the various outcomes mentioned above in the example of psychotherapy are aspects of mental health and can be lumped together for a general investigation of the effects of psychotherapy. When the effects are thought to perhaps vary among different outcome constructs, Glass suggests including data that indicate major distinctions among the constructs and using it as a predictor in multiple regression analyses or as a stratifying factor for nested analyses. Though Glass directs his suggestion to variations in the construct of the criterion, it is equally appropriate for variations in the construct of predictors.

The second aspect of the problem is one that past reviewers have often complained about. When different studies use different measures of the same construct, and when the measures have not been validated, there is a serious question about the equivalence of the values generated by the different measures. For some characteristics such as age and sex, there is seldom any problem, but for others such as self-image and social support, there often will be a problem for which there is no simple solution. It should be noted that it is incorrect to rationalize that what variations exist in the metric of some variable will only serve to reduce the strength of relationship between that variable and some second variable, and therefore can be ignored if strong relationships are found. This would be true if the variations in the metric are not correlated with the second variable, but generally there is no assurance that this will be true.

Glass and his students have already completed some work that reduces the third aspect of the problem. They have assembled equivalency functions for some statistics and developed others (Glass, 1977a, White, 1976). Some of these functions are mathematical identities, but others are approximations. To date that work has not indicated the conditions under which the approximations become poor ones. This is a fertile subject for future research.

A sixth problem faced in meta-analysis is that of achieving valid and reliable coding of the characteristics of the primary studies that are to be analyzed. This problem includes the previously discussed one, but extends beyond it. When the set of reviewed studies is relatively small, the coding is likely to be done by a single investigator. But coding, say, 40 studies, may require as many as 60 to 80 hours; and this work may be stretched over a 4- to 6-week period, thus raising serious threats to coding stability. When large numbers of studies are being reviewed, a number of coders may be used, which raises the additional problem of inter-coder reliability. Failures of memory, boredom, and migraine headaches can undermine sustained coding reliability. When the coding is done over a lengthy period of time, inter-coder reliability should be assessed more than once; reliability over time should also be assessed; and periodic retraining may be needed.

A seventh problem faced in meta-analysis is how to control for the effects of poor research design or execution among the reviewed studies. Glass (1976b) provocatively argued that it is wasteful to discard poorly designed studies from the analysis because, "a study with a half-dozen design or analysis flaws may be valid . . . [and] it is an empirical question whether relatively poorly designed studies give results significantly at variance with those of the best designed studies" (p. 4). Glass suggests testing whether methodological characteristics such as the reactivity of the outcome measure and the internal validity of the design are related to the distribution of findings. He does not specify how this should be done other than by examining the covariation between the design characteristics and the findings. The appropriate test, however, is not quite as straightforward as it may appear.

The relation can be examined in either a regression analysis or analysis of variance. Either model can yield misleading results under certain circumstances. Both models will be inadequate if there is not at least a modest number of studies with good overall internal and external validity. Since there is usually no reason to think that the relationship between validity and the findings is linear or monotonic, there is no basis for extrapolating from the relation that holds for studies of poor and mediocre validity to the studies with good validity. In integrative reviews of some topics, there may be very few if any primary studies with good internal and external validity, and thus in these cases the possible effects of less than good validity cannot be assessed. If there is a modest number of primary studies with good validity but a much larger number of studies with poor or mediocre validity, regression analyses of the full sample of cases can underestimate the effects of validity. This is because regression lines are fitted so as to minimize the squared deviations of the bivariate of multivariate points, and relatively little weight would be given to the small number of points from the good validity studies.

Both analysis of variance and regression analysis will underestimate the effects of validity if the mean level of the criterion is about the same for poor, medium and good validity studies, but the variance is considerably greater for the poor and medium validity studies. In such a case, both types of analysis will correctly indicate that varying validity does not affect the means of the criterion, but both would underestimate the effects of relatively poor validity on any correlations with the criterion. It should be noted that the variances of the different validity groups must be statistically significantly different for them to produce substantial underestimates of the effects of varying validity.

Earlier in this report it was indicated the congruence of findings does not assure their validity, and the lack of congruence is not proof of invalidity. When there is strong congruence in the findings and no good evidence of a strong common threat to the validity of all or almost all of the studies, there is suggestive, (but not conclusive), evidence that any methodological weakness that existed in some of the studies probably did not have a substantial effect on the findings of those studies. But in the more common situation when there are some apparent incongruences in the findings, it is important to have at least a modest number of studies in a sample that are judged to have good overall internal and external validity, if one is to empirically assess whether the methodological weaknesses that vary over the studies may have affected the findings.

What constitutes good enough overall internal and external validity for these purposes cannot be simply explicated. It probably depends on a number of factors, and needs further thought. It is important, however, to remember that threats of validity can be controlled by means other than design, and that some of the threats are likely to be trivial in any given study. This was discussed on page 10 of this paper.

Glass (1977b) has suggested that if the quality of a study is found to be related to the findings, a greater stake "should be put in the better designed study." This might be done by some sort of weighting, nesting analyses within subsets of the good and poor quality studies with more reliance put on the results of the former, or by disregarding the poor ones.

This discussion has very briefly outlined the major advantages of the use of meta-analysis for integrative reviews, and has presented, in considerably more detail, some difficulties with the approach. The disproportionate attention given to the difficulties should not mislead the reader into thinking that meta-analysis has more disadvantages

than advantages, or has more disadvantages than other analytical approaches when doing integrative reviews. In the opinion of this writer, the approach is methodologically sounder than most currently used approaches. Though it does have some serious difficulties, most of these difficulties are common to the other approaches. Also, the other approaches have additional difficulties or limitations which are not true of the meta-analysis.

In short, the meta-analytic approach is an important contribution to social science methodology. It is not a panacea, but it will often prove to be quite valuable when applied and interpreted with care.

Task 6: Interpreting the Results (V67, 68, 69, 70, 71)

Seven of the 36 randomly selected reviews induced and reported new theory, confirmation of old theory, or disconfirmation of old theory (V67); 6 of the 36 induced and stated recommendations for policy or practice, and four of those six discussed conditions which might affect the impact of the policies or practices (V68 & 69); 28 of the 36 suggested desirable foci or methods for future primary or secondary studies on the topic (V70); and only 3 of the 36 suggested desirable foci or methods for future reviews on the topic or related topics (V71).

There are other types of conclusions that the review articles may have stated that were not coded, but it is surprising that fewer than half made conclusions about either theory, policy or practice. It may be that most integrative reviews are oriented towards making suggestions for improving the primary research, or it may be that most start with the aim of making suggestions for theory, policy or practice, but subsequently decide to withhold inferences because they judge the available evidence to be inconclusive. The latter reason seems unlikely since the studied reviews usually did report one or more inferences about the topic. Of the 26 reviews that drew at least one inference of inconclusive evidence, 24 of those also drew at least one inference that an investigated condition or relation does exist either generally or for a specified subset of the population or of the investigated situations (V59a-d). (Multiple inferences were drawn in most of the reviews because they had multiple sub-topics that were investigated.)

This writer has no strong suspicions as to why most review articles do not make suggestions for future reviews. Perhaps it is because reviewers do not think carefully about the methods of doing a review; perhaps it is because after completing the often herculean task of doing a review, the reviewers would not want to wish the task on anyone else; or perhaps it is for some other reason. Regardless, it is quite apparent that the resulting omission is unnecessary and harmful to the progress of science. As with primary research, it is virtually impossible to do a major review carefully without encountering some ideas for improved

methods and some additional questions that need to be answered but cannot be answered in the given review. These ideas and questions can be a valuable contribution to other investigators and ought to be reported, even if they only can be used after the accumulation of further primary research.

Task 7: Reporting the Review (V7, 13, 17, 18, 44a-d)

A widely held precept in all the sciences is that reports of research ought to include enough information about the study that the reader can second guess the author's inferences. This precept probably also ought to apply to integrative reviews, since such reviews are a form of research. As a minimum, it is widely held that the report ought to at least describe the sampling, measurement, analyses, and the findings. Where unusual procedures have been used, it is expected that they will be described in some detail.

Some of the previously discussed results indicate that few of the 36 review articles reported certain methodological aspects of the review. Only one of the 36 articles reported whether or not it used indexes and information retrieval systems to search for primary studies on the topic (V16); 3 of the 36 reported whether or not they used bibliographies as a means of locating primary studies (V19); 7 of the 36 indicated whether or not they analyzed the full set of located studies on the topic instead of some subset (V45); and only one of the 36 indicated whether or not needed information that was missing in the reports of the primary studies was sought from other sources (V60).

A number of other aspects of a review that might be reported were coded. Thirty of the 36 articles explicitly stated the topic being reviewed (V7); 12 of the 27 articles that cited previous reviews indicated how their review was to differ from previous ones (V13); the one article that had indicated that it used indexes and information retrieval systems also indicated the beginning and ending dates and the descriptors that were used (V17 and 18).

Twenty-eight of the 36 review articles often reported the findings of an individual study or indicated how many or what percentage of the studies had each type of finding or result (V44a); three of the 36 articles often cited the range, mean or other summary indicator of the findings of the studies (V44b); and half of the 36 reviews often just reported a generalization followed by the citation of several studies (V44c); three approaches were coded independently and were not mutually exclusive.

These results, taken together, indicate that integrative review articles commonly fail to report their studies in the detail that is fairly common for primary research articles. A number of important functions are served by reporting various aspects of the review.

There are two reasons for carefully reporting the literature search process in an integrative review article. First it helps the reader to judge the comprehensiveness and representativeness of the sources that are the subject of the review. Just as the sample in a primary study can critically influence the findings of the study, the selection of the primary and secondary studies that are included in a review can seriously affect the results of the review. The bibliography of a review article indicates what individual studies were included in the review, but it does not indicate what broad classes of possibly relevant studies were excluded. A person with a thorough knowledge of the research on the topic will be able to infer such omissions by carefully examining the bibliography, but persons with less thorough knowledge of the topic will not be able to do so. Secondly, briefly detailing the literature search process in a review article allows future reviewers of the topic to extend easily the review without duplicating it. If it is known that most of the articles included in the review were those listed under certain descriptors of certain years of certain indexes, or found in the bibliographies of specified sources, it is very easy for a subsequent reviewer to broaden or deepen the search for relevant sources without duplicating the earlier work.

If some located primary studies were exluded from the analysis, the manner in which this was done ought to be reported. Likewise, how missing data are handled should be reported. An explicit statement of the topic being reviewed and an indication of how the reported review was to differ from previous ones on the topic often helps orient the reader and prevent misinterpretations.

When the number of reviewed studies is less than about forty, it is usually very easy to present a single page table indicating a number of the investigated characteristics of the primary studies including their findings, stated in either standardized or unstandardized form, or both, and with the direction and statistical significance indicated. Such information would allow any reader to reanalyze the studies and second guess the reviewer's analysis. Such opportunity is always a little threatening, but one of the oldest conventions of the scientific community is making one's data available to other scholars after one has had a chance to analyze and publish reports of it. When the number of primary studies is quite large, it is not practical to include all the data in the published report, but it should be available upon request (with adequate documentation).

The practice of reporting a generalization followed by the citation of several studies (V44c) was often used by half of the 36 reviews despite the fact that it can be very misleading. Unless used in conjunction with one of the other two approaches (V44a or V44b) this practice provides the reader with no way of critically examining the inferences of the reviewer unless he or she consults the full set of studies on the topic. For instance, Dusek (1975) reported "... there is considerable evidence that during classroom interactions teachers treat groups of students differently (e.g., Davidson, 1972; Good & Brophy, 1970; Schwebel & Cherlin, 1972)"

(p. 662). Hoffman (1972) reported, "Several investigators report that while dependency in boys is discouraged by parents, teachers, peers, and the mass media, it is more acceptable in girls (Kagan & Moss, 1962; Kagan, 1964; Sears, Rau, & Alpert, 1965)"(p. 144). Both of these statements give an implication of consensus among the available research evidence, but neither of the statements would be incorrect even if the majority of the located evidence contradicted their points. Jacoby made a similar type of statement, reporting "Not surprisingly studies which utilized price as the only independent variable (261, 262, 316, 452) generally found a significant main effect..." (p. 336). Though the "generally" in this statement provides an explicit warning that the evidence was not entirely consistent, it still does not indicate what percentage of the studies supported the finding, nor does it indicate the magnitude of the findings. Some reviewers stated juxtaposed generalizations such as, "Several studies found that X causes Y (Ace, 1967; Bace, 1953, Cace, 1969; Dace, 1970), but a few did not (Ease, 1968, Face, 1970)." This type of presentation is less ambiguous than the above examples, but this writer's impression is that it was not prevalent.

Other

A number of characteristics of the primary research that might have affected the methodological approaches used in each review were coded. These characteristics were as follows: whether the primary research was psychological, sociological, or about education (V2); whether the topic of the review was about a condition, association, or causal relation (V29 a-c); the types of construct investigated in the primary research (V30a-f); the predominant research orientation of the primary research (V34 and 72); the percentage of primary studies that investigated at least one statistical interaction effect (V36); and a crude estimate of when research was first done on the given topic (V37b).

Most of the characteristics of the primary research were cross-tabulated with some of the variables indicating the different methodological approaches of the reviews. The variables that were excluded from these analyses were those whose distribution in their original form or a conceptually reasonable transformation would frequently have resulted in expected cell sizes that would make a Chi-square test of the cross-tabulation invalid. Thus V2, 75, 76, 77, 78 and 78 were each cross-tabulated with each V11, 41, 44a, 44c, 53a, 56c.

Only one of these 36 cross-tabulations had a Chi-square statistically significant at the 0.05 level or less, but the Chi-square for this cross-tabulation was invalid because more than 20 percent of the cells had expected values of less than 5. In addition, when testing 36 hypotheses at the 0.05 level, one or two false rejections of the null hypothesis can be expected from chance if the hypothesis is true. Consequently, the analysis failed to discover reliable evidence of a relationship between any of the examined characteristics of the reviewed research and any of the examined approaches to the methodological tasks of an integrative review. This, of course, should not be interpreted as indicating that there are no such relationships, but only that given the small sample size and the skewed distributions of some of the variables of interest, no reliable inferences could be made.

Results and Discussion of Other Sources of Information on Integrative Review Methods

Only five published rejoinders to the 36 randomly sampled reviews could be located. The analysis of these rejoinders was not terribly enlightening and will not be discussed in this paper.

The editors of five prestigious social science journals that frequently publish integrative reviews were asked about the evaluation criteria and standards that they use to decide whether or not to publish submitted integrative reviews. One editor provided printed guidelines that he provides to prospective reviewers and his editorial assistants, one editor failed to reply to repeated followups, and three editors said essentially that they rely on the professional judgments of their editorial assistants or authors of invited reviews.

The officials of ten organizations that were thought to have major responsibilities for integrative reviews in the social, biological and physical sciences were asked about: 1) the formal or informal guidelines or standards used by their offices to facilitate high-quality reviews of sets of empirical research studies, 2) the evaluative criteria used to judge the quality of such reviews, and 3) examples of such reviews that they consider to be of unusually high quality. There were basically three types of responses. A couple of the respondents reported some guidelines or evaluative standards, but they generally were not very specific. Some of the respondents indicated that they rely almost exclusively on the judgments of the scientists who they have do their reviews. And two of the respondents thought that integrative reviews were not often done in their disciplines (math, physics and space sciences), though subsequently it was discovered by this author that the Reviews of Modern Physics frequently published such reviews.

Conclusion

It appears that relatively little thought has been given to the methods for doing integrative reviews. It is clear that such reviews are important to science and social policy-making and that many integrative reviews are done less rigorously than is currently possible. It seems likely that some of the confusion that surrounds many topics in the social sciences is partly a result of unrigorous reviews of research on the topic.

This paper and the recent work of Gene Glass provide several ideas for improving the prevailing methods for reviews. None of the ideas should at this stage be considered definitive. Rather, there is need for scientists who do or use integrative reviews to consider the merits of the ideas, to think more about the problems to which they are directed, to try new approaches that appear promising, and to evaluate the effectiveness of those approaches.

References

Barnes, B. R. & Clawson, E. V. Do advance organizers facilitate learning? Recommendations for further research based on an analysis of 32 studies. Review of Educational Research, 1975, 45, 637-659.

Bracht, G. H., & Glass, G. V. The external validity of experiments. American Educational Research Journal, 1968, 5, 437-469.

Bradley, J. V. Distribution-free statistical tests. Englewood Cliffs, N.J.: Prentice-Hall, 1968.

Campbell, D. T., & Stanley, J. C. Experimental and quasi-experimental designs for research. Chicago: Rand McNally, 1963.

Coleman, J. S. Recent trends in school integration. Educational Researcher, 1975, 4(7), 3-12.

Demerath III, N. J., & Roof, W. C. Relition--recent strands in research, Annual Review of Sociology, 1976, 2, 19-33.

Dusek, J. B. Do teachers bias children's learning?, Review of Educational Research, Fall 1975, 45(4), 661-684.

Feldman, K. A. Using the work of others; some observation on reviewing and integrating. Sociology of Education, 1971, 44, 86-102.

Glass, G. V. Primary, secondary, and meta-analysis of research. Presidential address to the Annual Meeting of the American Educational Research Association, San Francisco, April 21, 1976(a).

Glass, G. V. Primary, secondary, and meta-analysis of research. Educational Researcher, 1976(b), 5, 3-8.

Glass, G. V. et al. Teacher 'indirectness' and pupil achievement: An integration of findings. Unpublished manuscript, University of Colorado, 1977(a).

Glass, G. V. Personal communication, October, 1977(b).

Glass, G. V. Integrating findings; The meta-analysis of research. Review of Research in Education, 1977(c), 5,

Glass, G. V. Personal communication (notes on margins of draft report that he read), November 1977(d).

Hays, W. L. Statistics. New York: Holt, Rinehart & Winston, 1963.

Hoffman, L. W. Early childhood experiences and women's achievement motives. Journal of Social Issues, 1972, 28(2), 129-155.

Jacoby, J. Consumer psychology: An octennium. Annual Review of Psychology, 1976, 27, 331-358.

Kerlinger, F. N., & Pedhazur, E. J. Multiple regression in behavioral research. New York: Holt, Rinehart & Winston, 1973.

Kish, L. Survey sampling. New York: John Wiley, 1965.

Lambert, M. J. Spontaneous remission in adult neurotic disorders: A revision and summary. Psychological Bulletin, 1976, 83(1), 107-119.

Light, R. J., & smith, P. V. Accumulating evidence: Procedures for resolving contradictions among different research studies. Harvard Educational Review, 1971, 41, 429-471.

Platt, J. R. Strong inference. Science, 1964, 146, 347-353.

Schultz, C. B., & Sherman, R. H. Social class, development, and differences in reinforcer effectiveness. Review of Education Research, 1976, 46, 25-59.

Sechrest, L. Personality. Annual Review of Psychology, 1976, 27, 1-27.

White, K. R. The relationship between socioeconomic status and academic achievement. (Doctoral dissertation, University of Colorado, 1976), Dissertation Abstracts International, 1977, 38, 5067A-5068-A. University Microfilms No. 77-3250.)